Real-Valued Jaccard and Coincidence Based Hierarchical Clustering

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

27rd Oct 2021

Abstract

Hierarchical clustering represents one of the most frequently adopted methodologies for identifying clusters in data in non-supervised classification tasks. Amongst the advantages of this family of approaches, we have that the possible solutions are obtained in a multiscale manner involving a respective dendrogram of the data. In addition to providing a more complete description of the interrelationships between the data elements, the number of clusters does not need to be specified as in other clustering methods such as k-means, as it can be inferred from the obtained dendrograms. There are several possible hierarchical clustering methods, depending on the adopted merging criterion, which can be the smallest distance between sets (single linkage), or the minimization of dispersion (Ward's). The Jaccard index has also be considered for binary data. In this work, we propose a new family of hierarchical clustering methods, based on recent developments in which the Jaccard index is generalized to real values as well as on the coincidence index, which corresponds to the product between this generalized index and the interiority (or homogeneity) index. The former of these indices is more discriminative of anti-correlations, and the latter also provides a more strict comparison of the involved clusters. Therefore, it is expected that the coincidence index-based hierarchical cluster be less likely to yield false positive clusters than other hierarchical approaches. In addition, it becomes possible to start with the elements to be clustered represented by generic densities or even general scalar fields.

'The old tree, merged into the ground and sky.'

LdaFC

1 Introduction

Hierarchical clustering (e.g. [1, 2, 3]) constitutes one of the most often employed non-supervised methods as a consequence of its interesting features. Though there are two main classes of hierarchical clustering methods, namely divisive and agglomerative, in the present work we will focus on the latter.

Basically, agglomerative hierarchical clustering methods are characterized by the subsequent merging of clusters, defining a respective dendrogram that provides valuable information about the interrelationships between the data and subclusters. In addition, unlike other methods such as k-means, the number of clusters does not need to be pre-specified. Actually, the most likely number can often be estimated while taking into account the structure of the obtained dendrogram. In addition, agglomerative methods are conceptual and computationally simple.

Agglomerative hierarchical clustering are defined respectively to the criterion adopted for merging the clusters. For instance, in the single-linkage approach, the merging proceeds based on the smallest distance between the existing clusters. In the extensively used Ward's method, the merging aims at maintaining the smallest dispersion of the clusters. Therefore, a virtually an infinite number of possible agglomerative methods because there is an infinite number of possible merging criterion.

Though potentially interesting, the use of the Jaccard similarity index in hierarchical clustering has been mostly limited to comparing sets (e.g. [4]). However, this index can be generalized to take into account real data [5, 6, 7], including respective densities.

In particular, it has been shown [7] that the generalized Jaccard index relates directly to the prototypical similarity quantification by using the Kronecker's delta function. More specifically, the Jaccard can be understood as a respective version that is more tolerant than the absolutely strict Kronecker-base criterion regarding data similarity. At the same time, the generalized Jaccard has also been found [7] to be more robust than the cosine distance as it penalizes more intensely the existing anti-correlations.

Given that the classic Jaccard similarity index does not take into account the relative interiority between the com-

pared sets, a new respective generalization has been proposed that incorporates an additional index for quantification of the interiority (or homgeneity) of the pairwise set combinations [5, 6, 7]. The resulting similarity index therefore provides a more strict quantification of the similarity between sets, vectors or even real functions.

In the present work, we develop two new types of agglomerative hierarchical clustering based respectively on the real-valued Jaccard and coincidence inidices.

We start by reviewing the two adopted index and then present the respectively obtained agglomerative hierarchical clustering methodologies. A simple example is also provided in order to illustrate the proposed approaches.

2 The Jaccard Index for Real Values and the Coincidence Index

The traditional Jaccard index [5]) has been extensively employed as a measurement of the similarity between two sets A and B, being defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

This index has been recently generalized to take into account multisets with possibly negative, real values [5, 6, 7] as:

$$s_1(A,B) = \frac{\sum_{i \in S} s_{x_i y_i} \min\{s_{x_i} x_i, s_{y_i} y_i\}}{\sum_{i \in S} \max\{s_{x_i} x_i, s_{y_i} y_i\}} \tag{2}$$

where the multiplicities of the sets A and B are represented as x_i and y_i, respectively, and S is the shared multiset support, i.e. the elements underlying both multisets (e.g. [6]).

The Jaccard index has been shown not be able to take into account the relative interiority of the two sets [5]. However, we can adopt the following interiority (or homogeneity) index:

$$I(A,B) = \frac{\sum_{i \in S} \min\{s_{x_i} x_i, s_{y_i} y_i\}}{\min\left\{\sum_{i \in S} s_{x_i} x_i,\ \sum_{i \in S} s_{y_i} y_i\right\}} \tag{3}$$

$$\tag{4}$$

So that the following new index, namely the *coincidence index*, can be obtained:

$$C(A,B) = I(A,B) J(A,B) \tag{5}$$

3 Jaccard and Coincidence-Based Hierarchical Clustering

The real-valued Jaccard index, as well as the coincidence index, can be adopted in order to obtain two respective new types of agglomerative hierarchical clustering.

The basic idea is to proceed with the cluster merging so that the two current clusters presenting the largest similarity, as quantified by the real-valued Jaccard or coincidence indices, are merged at each step.

Because the two similarity indices compare densities, kernel expansion of the current clusters are required at each step. Though in the current work we adopt circularly symmetric gaussian kernels, other generic choices can be adopted to suit specific requirements. In particular, it becomes possible to start with the individual elements representing not only isolated points, but whole densities of generic types and shapes.

So, the real-valued and coincidence agglomerative clustering methods can be summarized as:

Input data elements $c_1, c_2, \ldots, c_N$;
$m = 1$;
While$(m < N)$:
Perform kernel expansion of the current clusters;
Calculate the indices between these clusters;
Join the two clusters that are more similar;
Save list of obtained clusters;
$m = m + 1$;

Observe that the calculation of the densities for the merged clusters can be immediately obtained by summing the respectively involved densities, without need of additional kernel expansions. The densities are assumed to be normalized in the sense of having unit area.

4 Case Example

As a simple case example, consider the distribution of individuals represented in terms of two respective features as depicted in Figure 1.

The successive mergings obtained by the coincidence-based hirarchical clustering is as follows:

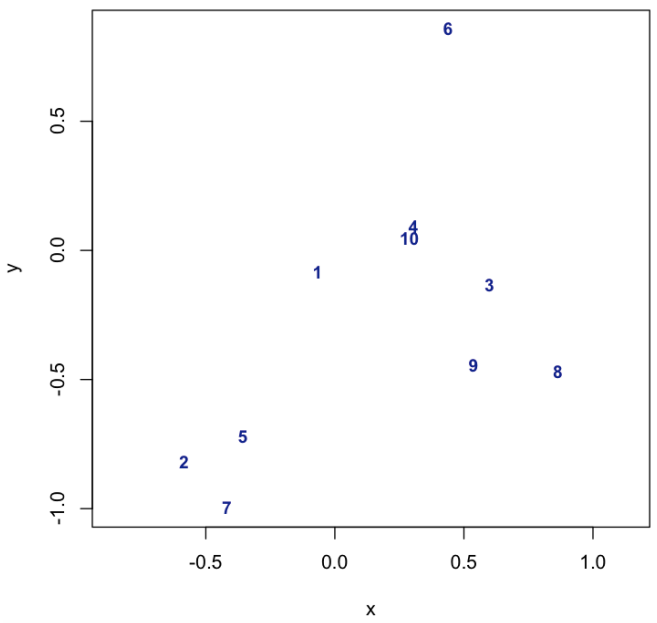

Figure 1: Scatterplot for the case example.

$$
\begin{aligned}
m = 1: &\quad 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \\
m = 2: &\quad 1, 2, 3, (4, 10), 5, 6, 7, 8, 9 \\
m = 3: &\quad 1, (2, 7), 3, (4, 10), 5, 6, 8, 9 \\
m = 4: &\quad 1, (2, 7, 5), 3, (4, 10), 6, 8, 9 \\
m = 5: &\quad 1, (2, 7, 5), (3, 9), (4, 10), 6, 8 \\
m = 6: &\quad 1, (2, 7, 5), (3, 9, 8), (4, 10), 6 \\
m = 7: &\quad (1, 4, 10), (2, 7, 5), (3, 9, 8), 6 \\
m = 8: &\quad (1, 4, 10, 3, 9, 8), (2, 7, 5), 6 \\
m = 9: &\quad (1, 4, 10, 3, 9, 8, 6), (2, 7, 5) \\
m = 10: &\quad (1, 4, 10, 3, 9, 8, 6, 2, 7, 5)
\end{aligned}
\tag{6}
$$

5 Concluding Remarks

Agglomerative hierarchical clustering has been extensively used in a large number of scientific and technological areas (e.g. [1, 8, 2, 3]). Several of the most traditional types of these methods have relied on distances between the clusters, while the the Ward's approach consists of merging clusters so as to ensure minimal dispersion. Though potentially interesting, the Jaccard index has been mostly constrained to agglomerative clustering applications involving quantifications of similarity between sets of objects.

Two new types of agglomerative hierarchical clustering methods have been proposed, respectively based on the real-valued Jaccard and coincidence indices [5, 7]. These two generalizations of the classic Jaccard similarity index have been proposed recently to cope with negative real data values, while the latter index also takes into account the relative interiority between the involved sets [5, 6, 7]. These two generalizations have relied on data representation as multisets (e.g. [9, 10, 11, 12, 13, 14]), understanding the multiplicity to encompass real values, including possibly negative quantities.

Several interesting features are provided by the proposed methodology to hierarchical clustering. First, we have that the original data elements can be not only individual observation, but respective discrete (or even continuous) generic density distributions, which can have any type of shape. Second, it becomes possible to compare not only non-negative densities, but any scalar fields associated to the data elements, including negative values. Third, we have that the Jaccard and coincidence indices have been shown to be more robust for cluster comparison because the impose a higher penalty on anti-correlations between the involved distributions. The coincidence index, in particular, takes into account the relative interiority of the densities, therefore implementing a more strict comparison between the densities. The latter feature is of particular potential relevance, because clustering methods have been shown to present a tendency to false positive identification of clusters [3].

Several further studies are motivated by the methodology proposed in this work, including systematic comparisons between several types of hierarchical (as well as other types) of clustering approaches, the consideration of different types of kernes possibly reflecting specific application requirements, as well as evaluating the methodologies for higher dimensional feature spaces.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[2] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[3] E. K. Tokuda, C. H. Comin, and L. da F. Costa. Revisiting agglomerative clustering. *Physica A*, 585:126433, 2022.

[4] Doruk Kilitcioglu. Hierarchical clustering and its applications. https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6, 2018. [Online; accessed 30-Oct-2021].

[5] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[6] L. da F. Costa. Multisets. https://www.researchgate.net/publication/355437006_Multisets, 2021. [Online; accessed 21-Aug-2021].

[7] L. da F. Costa. On similarity. https://www.researchgate.net/publication/355792673_On_Similarity, 2021. [Online; accessed 21-Aug-2021].

[8] L. da F. Costa and R. M. C. Cesar Jr. *Shape Classification and Analysis: Theory and Practice*. CRC Press, Boca Raton, 2nd edition, 2009.

[9] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[10] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[11] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[12] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[13] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[14] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.